Cheaper Faster Drug Development Validated by the Repositioning of Drugs Against Neglected Tropical Diseases

Kevin Williams[1], Elizabeth Bilsland[2], Andrew Sparkes[1], Wayne Aubrey[1], Michael Young[3], Larisa N. Soldatova[4], Kurt De Grave[5], Jan Ramon[5], Michaela de Clare[2], Worachart Sirawaraporn[6], Stephen G. Oliver[2], Ross D. King[7]

1. Department of Computer Science, Aberystwyth University, SY23 3DB, UK
2. Cambridge Systems Biology Centre & Department of Biochemistry, University of Cambridge, Sanger Building, 80 Tennis Court Road, Cambridge CB2 1GA, UK
3. Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, SY23 3DD, UK
4. School of Computer Science, Brunel University, UB8 3PH, UK
5. Department of Computer Science, KU Leuven, 3001 Heverlee, Belgium.
6. Department of Biochemistry, Mahidol University, Thailand.
7. Manchester Institute of Biotechnology & School of Computer Science, University of Manchester, Manchester, M1 7DN, UK.

Abstract:
There is an urgent need to make drug discovery cheaper and faster. This will enable the development of treatments for diseases currently neglected for economic reasons, such as tropical and orphan diseases, and generally increase the supply of new drugs. Here we report the Robot Scientist "Eve" designed to make drug discovery more economical. A Robot Scientist is a laboratory automation system that uses artificial intelligence (AI) techniques to discover scientific knowledge through cycles of experimentation. Eve integrates and automates library screening, hit confirmation, and lead generation through Quantitative Structure Activity Relationship (QSAR) learning and testing. Using econometric modelling we demonstrate that the use of AI to select compounds economically outperforms standard drug screening. For further efficiency Eve utilizes a standardized form of assay to compute Boolean functions of compound properties. These assays can be quickly and cheaply engineered using synthetic biology, enabling more targets to be assayed for a given budget. Eve has repositioned several drugs against specific targets in parasites that cause tropical diseases. One validated discovery is that the anti-cancer compound TNP-470 is a potent inhibitor of dihydrofolate reductase in the malaria-causing parasite *Plasmodium vivax*.

1. Introduction

New drugs are generally slow (> 10 years) and expensive (> $1 Billion) to discover and develop. Consequently tropical diseases, malaria, schistosomiasis, Chagas' disease, etc., which kill millions of people and infect hundreds of millions of others are "neglected" (Ioset & Chang, 2011; Leslie, 2011); and "orphan" diseases with few sufferers remain untreatable (Braum *et al*., 2010). More generally, the pharmaceutical industry is struggling to cope with spiraling drug discovery and development costs (Pammolli *et al*., 2011).

The most important steps in early-stage drug design are shown in Fig. 1 (Gad, 2005). A key initial step is to develop an 'assay'. This is a 'wet' (biological/chemical) or 'dry' (computational) experiment that estimates whether a small molecule (compound) is likely to treat a disease. This assay should be relatively cheap and fast to execute, as it will be run multiple times. A compound that passes the assay is called a 'hit'. The next step is to run a drug screen, where a "library" (set) of compounds is tested against the assay. This library may be very large, tens/hundreds of thousands, maybe millions of compounds. Such mass screening is generally done in a brute-force and unintelligent way: 'begin at the beginning and go on till you come to the end: then stop' (Lewis Carroll). As the *a priori* probability of any library compound being a hit is low, it is difficult to design an assay that does not have an appreciable number of false positive hits. Therefore, it is generally necessary to execute experiments to retest ('confirm') the hits. These experiments are more expensive and slow to execute, but have a far lower false positive probability. From the set of confirmed hit activities a quantitative structure activity relationship (QSAR) is learnt (Martin, 2010). This is a function whose input is the structure of a compound, and whose output is the predicted activity on the assay. As the output is typically a real number, QSAR learning is generally a regression task. QSARs generalise the results of assays and guide the synthesis of new compounds. After new compounds are synthesised they are tested against the hit confirmation assay, and the results of these assays are used to learn a more accurate QSAR, and the cycle repeated. The process is terminated when a compound is found that that has a sufficiently high score on the assay, and passes other tests such as low predicted toxicity, potential for modification, etc. This compound is called a 'lead'.

The standard way to improve the economics of a process is through automation and standardization (Bernal, 1969). The use of automation has been enthusiastically pursued by the pharmaceutical industry. With much of this effort has gone into making library screening faster, especially through miniaturisation, with the result that high-throughput robotic systems now routinely screen millions of compounds in library screens (Gad, 2005). Less automation effort has gone into the other steps of early stage drug design. Standardization has been little used in early stage drug design.

A natural extension of the trend of increased involvement of automation in science is the concept of a Robot Scientist (King *et al.*, 2004; King *et al.*, 2009). A Robot Scientist automatically: originates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments using laboratory robotics, interprets the results to change the probability of hypotheses, and then repeats the cycle. In this way Robot Scientists can automate high-throughput hypothesis led research. Robot Scientists are also well suited to recording scientific knowledge: as the experiments are conceived and executed automatically by computer, it is possible to completely capture and digitally curate all aspects of the scientific process. (King *et al.*, 2009). The first Robot Scientist 'Adam' was designed to plan and execute yeast microbiological experiments. Adam was fully automated and there is no essential requirement for a technician, except to periodically add laboratory consumables and remove waste. Adam was the first machine demonstrated to have autonomously discovered novel scientific knowledge (King *et al.*, 2009). Adam investigated the functional genomics of *S. cerevisiae* and discovered the function of locally orphan enzymes – enzymes known to be in yeast but for which the gene(s) encoding them are unknown (King et al., 2009). The advances that distinguished Adam from other complex laboratory systems (such as high-throughput drug-screening pipelines, and X-ray crystallography crystal-screening systems) was its AI software, its

many complex internal cycles, and its ability to execute individually planned cycles of experiments in high-throughput.

In this paper we concentrate on finding lead compounds for neglected tropical diseases, however we believe the methods used and conclusions are general. The reasons for the focus on neglected tropical diseases are:

- These disease are a scourge of humanity, infecting hundreds of millions of people, and annually killing millions of people.
- The aetiology of these diseases is clear, as is what needs to be done to treat the disease (kill the parasites), and how to achieve this treatment with a small molecule drug. These are not the case for many diseases targeted by the Pharmaceutical Industry.
- There is little competition from the much better funded Pharmaceutical Industry.

2. Eve

2.1 Design

We report the development of the Robot Scientist "Eve" designed to automate early-stage drug design (Fig 2). The initial design of Eve was given in (Sparkes *et al*. 2010). Eve has three integrated modes corresponding to successive stages in lead drug discovery. In its Library-screening mode Eve systematically tests each member from a large set "library" of compounds against an assay in the standard brute-force way of conventional mass screening (Gad, 2005): While simple to automate, brute-force mass screening is slow and wasteful of resources as every compound in the library is tested. It is also unintelligent, as it makes no use of what is learnt during screening. Eve starts the lead discovery process by mass-screening a subset of its library to find "hit" compounds for the assay. This subset was chosen randomly.

In its Hit-confirmation mode Eve re-assays the hit compounds using multiple repeats and titrations to reduce the probability of false-positives. Eve's integration of screening and hit confirmation is similar to advanced screening systems that first execute a high-throughput screen, and then a high-content screen for selected compounds.

Starting from the set of confirmed hits, Eve executes cycles of statistics/machine learning that hypothesize QSARs, and tests these QSARs on new compounds. As Eve currently does not have access to chemical synthesis automation (Baumann *et al*., 2011), we applied Eve to screen untested compounds from its library in lieu of synthesizing compounds. Such intelligent library screening may be more economical than standard mass screening as it potentially saves on time and compound use.

2.2 Hardware

Eve's robotic system is capable of moderately high-throughput compound screening: greater than 10,000 compounds per day, depending on the length of time taken to assay compounds. Eve is designed to be sufficiently flexible that it can be rapidly re-configured to carry out a number of different biological assays, using fluorescence, absorbance, or cell morphology as read-outs (Fig. 2). Eve's robotic system integrates a range of off-the-shelf pieces of laboratory automation equipment into a single system that can perform library- screening, hit confirmation, and cycles of QSAR hypothesis formation and testing using selected compounds from a compound library. Eve can also be re-configured to copy compound libraries.

Eve's compound library is maintained in a dry-store, with the compounds dissolved in DMSO. Compounds to be assayed are transferred from the storage plates to the assay plates using a non-contact acoustic transfer liquid-handling system. This has the advantages of high accuracy and saving pipette tips. In library-screening mode, there is a direct mapping from storage plates to assay plates, and a single transfer volume is used in each well. In confirmation and intelligent-screening modes, a single compound from the storage plate is transferred to multiple wells in the assay plate, and at different volumes, to realize multiple repeats dose-response experiments. After the addition of assay compound, the target yeast strain pool is added using a simple liquid-handling robot, as the same volume is added to each well. The yeast pool is created externally and stored by Eve for use. Once the assay plates are formed they are placed in a shaking incubator. Every 90 minutes, the plates are removed from the incubator, and fluorescence measured. Eve has two microplate readers capable of recording measurements across a broad range of both excitation and emission wavelengths. Eve also has an automated microscope capable of taking both bright-field and fluorescence images across a broad range of wavelengths. Upon completion of the assay, the plates are automatically removed from the system. To transfer the plates between different pieces of laboratory automation equipment, Eve uses robotic arms and linear actuators. All plates are bar-coded and movements recorded.

2.3. Low-level Software

Software was written to integrate Eve's AI software with the robotics and thereby automate and integrate Eve's early-stage drug-design functions: library-screening, hit confirmation, and QSAR cycles. The software to control the robotics, instrumentation, and used to execute the experiments was written on top of PAA's (PAA Peak Analysis and Automation Ltd, Farnborough, UK) Overlord software. An interface was written to a relational database that stores all experiment-related data and meta-data, e.g. all fluorescence measurements. The software to parameterise growth curves for the different yeast strains in each well was taken from the Adam project (King *et al.*, 2009). The main parameters are maximum growth, doubling-time, and lag-time. These growth parameters were then transferred to the AI QSAR software. Software was also written to coordinate library-screening, and to plan hit confirmation. This was also integrated with the AI software so that the active learning could select compounds to test.

2.4. Automated QSAR Formation

To form QSARs Eve uses least-squares linear regression with mild 2-norm regularization (ridge regression). This can be interpreted as a Gaussian process with a linear kernel (Rasmussen & Williams, 2006), hence we can compute the posterior uncertainty, allowing us to use an optimization method which is more efficient, i.e. which needs fewer function evaluations (De Grave *et al.*, 2008). The linear kernel choice has the distinct advantage that it permits more efficient computation than other kernels when the feature space dimensionality is smaller than the number of examples. The feature space consists of binary fingerprints of all paths up to length 7. We computed these with Open Babel (O'Boyle, *et al*., 2011).

2.5. Active Learning

To select compounds to test its hypotheses Eve uses active learning (Cohen *et al*., 1996); De Grave *et al.*, 2008). The active learning task is comparable to that in many other areas of science and engineering: identify or design artifacts that have optimal performance. However, it has an extra ingredient reminiscent of reinforcement learning: balancing the *exploration* of compound

space with *exploitation* of regions of highly active compounds. Another complication is that it is desirable to identify the K best diverse compounds in the library: "leads" (De Grave *et al.*, 2008). Therefore, the QSAR active learning problem is: given a finite pool P of instances, an unknown QSAR function f that maps instances $x \in P$ to their target values $f(x)$, an assay (noisy "oracle") that can be queried for the target value of any x in batches of N, the number K of leads required; then find the top K leads in P. (In computer science an 'oracle' is a machine, the working of which are unexplained, which always returns the correct answer to a question. A noisy oracle has a probability of returning an incorrect answer.) We found a successful approach to be a combination of selecting compounds with high estimated activity T, and high estimated variance, i.e. select the example where $T + b\sqrt{var(T)}$ is maximal (De Grave *et al.*, 2008) (Supplementary material). As it is generally inefficient to assay (or synthesize) a single compound in a QSAR cycle, batches of N compounds should be selected (for Eve, $N=64$). Like the requirement to find the K best leads this greatly increases the computational complexity of choosing the best experiment. Therefore Eve adopts a greedy strategy to select batch compounds.

3. Standardized Assays

There are three main forms of assay: computational (Lounkine *et al.*, 2012), biochemical, and cell-based (Vogel, 2002). The most general type is computational (*in silico* screening, Schneider, 2010) - assuming the Church-Turing thesis (Deutsch, 1985), they could compute any conceivable assay. The advantages of *in silico* screening are that it is cheap, fast, and that compounds can be tested without synthesizing them. These enable very large libraries to be evaluated, and *in silico* screening has proved its worth many time (Schneider, 2010). The main disadvantage of *in silico* screening is that it is still computationally infeasible to simulate the full complexity of biological systems. Biochemical assays have the advantage of being target-based (enabling rational drug design), but often assume a specific mechanism of interaction, and provide little information about toxicity, drug uptake into cells, and *in vivo* activity. Cell-based assays are the most biologically realistic, but are rarely target based, and thus provide limited information on the mechanism of action of a drug. Moreover, cell-based assays are not applicable when searching for compounds active against parasites that are not currently possible or difficult to culture (e.g. *Plasmodium vivax*). All these types of assay are slow and expensive to develop - even computational ones if reasonable realism is to be achieved.

We have developed a standardized form of screening assay that combines advantages of computational assays (generality), biochemical assays (targeted), and utilizing live cells (biological realism, and early screening for toxicity) (Fig 3). These assays are designed to be automatically engineered using existing laboratory automation, and can be generated much faster and more cheaply than the bespoke assays that are currently standard. This enables more types of assay to be executed, more efficient use of screening facilities to be made, and thereby increases the probability of a discovery within a given budget. The assays are biological systems designed to compute Boolean functions of desired properties (Bememson *et al.*, 2004). This concept generalizes previous uses of engineered cells in drug discovery assays (Smith *et al.*, 2010; Bilsland *et al.*, 2011; Bilsland *et al.*, 2013). As an illustration, consider the example of designing an assay that targets both wild type and drug-resistant DHFR against *P. vivax* (Figure 3). To compute this function we first engineer a chimeric yeast (*Saccharomyces cerevisiae*) strain with its DHFR coding sequence (cds) replaced by that for wild type *P. vivax* DHFR

(y^{Pv}DHFRp**)**, then engineer a second chimeric yeast strain (y^{Pv}Rdhfrp**)** with its DHFR cds replaced by that for drug-resistant *P. vivax* DHFR. We then engineer a third chimeric yeast strain (y^{Hs}DHFRp**)** with its DHFR cds replaced by that for *H. sapiens* DHFR. Finally, we apply the biological system to assay for compounds that inhibit growth of the strains expressing the parasite targets (y^{Pv}DHFRp and y^{Pv}Rdhfrp) and not the strain expressing their human counterpart (y^{Hs}DHFRp) (Chong & Sullivan, 2007). Such compounds are 'true' for the assay. They are unlikely to be cytotoxic as one would expect all three strains to be inhibited. (Of course this does not completely remove the probability of human cytotoxicity as there could be off target effects specific to human cells, therefore further studies may be required). In practice Eve grows the strains in competition mixed cultures in 384-well microtitre plates (Bilsland *et al.*, 2013) in the presence of one compound from its library. The whole system is a *model* of what we really are interested in: the *in vivo* survival of wild type/drug-resistant *P. vivax* cells versus those of its human host. It can easily be seen that a set of genetically engineered yeast strains can compute arbitrary complex Boolean functions of desired assay properties.

4 Drug Screening and Repositioning

4.1. Standardised Assays

We first demonstrated that we could efficiently generate standardized assays. We generated assays targeting DHFRs (wild type and drug resistant), N-myristoyltransferase (NMT), and phosphoglycerate kinase (PGK) from multiple parasitic organisms:, *Trypanosoma brucei* (African sleeping sickness), *T. cruzi* (Chagas disease), *Leishmania major* (Leishmaniasis), and *Schistosoma mansoni* (Schistosomiasis) (Supplementary material). These assays were much faster and cheaper to develop than using standard methods of assay development: engineering each assay took about one person-month, and cost ~$15k. A subset of these assays were reported in Bilsland *et al*., 2013.

4.2. Drug Screening

We then tested the utility of these assays, and the efficiency of Eve at standard screening, i.e. running in its Library-screening and Hit-confirmation modes (Table 1). We ran the Maybridge Hitfinder library of approximately 14,400 chemically diverse compounds to these assays. This identified numerous hits. A subset of these results were reported in Bilsland *et al*., 2013.

4.3. Drug Screening for Drug Repositioning

We then applied the assays to the challenge of drug repositioning – the application of known drugs to new diseases (Table 1). To do this we again utilized Eve in its Library-screening and Hit-confirmation modes to screen and confirm hits for the above assays, but using the Johns Hopkins University Clinical Compound Library that contains ~1,600 FDA-and foreign-approved drugs. Several repositioned compounds were found that discriminate between host and parasite, and have passed initial cytotoxicity tests. To maximize the utility and reuse of these screening data they are available as open data in Resource Description Framework (RDF) format (Bizer *et al.*, 2009) at: http://disc.brunel.ac.uk/eve-dataset/RDF_data_files.html with an associated ontology (Supplementary material).

4.4. Repositioning TNP-470 as an Anti-malaria Compound

The compound TNP-470 was derived from the antimicrobial compound fumaglillin (Fig 4), it is an angiogenesis inhibitor that has been investigated as an anti-cancer drug. Its angiogenesis activity is mediated by irreversible binding to and inactivating of methionine aminopeptidase-2 (MetAP2). TNP-470 and its analogs have been shown to bind to *P. falciparum* MetAP2 *in vitro*, to inhibit growth of *P. falciparum* strains (including the chloroquine resistant strains W2 and C2B*)*, and to inhibit parasitemia in a mouse model (Zhang *et al.*, 2002; Chen *et al*., 2009; Arico-Muendel *et al.*, 2009). Eve's yeast synthetic biology assay results indicate that TNP-470 has high activity against *P. vivax* DHFR (Fig 5). To further confirm that DHFR is an additional target of TNP-470, we performed DHFR enzyme inhibition assays (Leartsakulpanich *et al*., 2002). We observed that *P.vivax* DHFR was 100-fold more sensitive to TNP-40 than its human counterpart; the drug's IC_{50} for the parasite enzyme being 0.16 µM, compared to > 165 µM for human DHFR. This is consistent with the results of Eve's assays, and suggests that our approach identified a *bona fide* DHFR inhibitor with improved selectivity.

DHFR inhibitors are currently routinely used as prophylactics against malaria, and are given to over a million children in seasonal malaria chemoprevention. However, DHFR inhibitors are no longer used as a standard treatment because of the evolution of drug resistance (22). Extensive efforts to discover a second-generation DHFR-targeted antimalarial drug with efficacy against pyrimethamine-resistant strains have yet to produce a compound that has passed clinical trials (Yuthavong *et al.* 2012). Therefore, the discovery of an approved compound with activity against DHFR is of high potential value. It is also significant that TNP-470 is an example of "polypharmacology" (Besnard *et al.*, 2012), in that it targets both *Plasmodium* DHFR and MetAP2. This means that it should be pre-hardened to the evolution of drug resistance, as this would require simultaneous alteration of both targets.

5. Automating Drug Development

5.1 Automating Drug Development

We integrated all three of Eve's modes together to demonstrate that early-stage drug development can be automated, including QSAR generate-and-test cycles. The division of labour between Eve and the human scientists and technicians was as follows: the problem task was first tightly defined by the humans by forming the assays, and defining the QSAR problem. This was the extent of human intellectual effort. Human manual effort was required to maintain run Eve, maintain consumables, yeast stocks, etc. Human manual effort was also required to run certain programs during the different stages of the cycles, as some of the steps are not fully integrated; these program steps are predetermined, and could if necessary be fully automated.

The first full experimental tests of the active learning loop were conducted by splitting the screen data set for TS6 (comprising the heterologous DHFR yeast strains for *P. falciparum, P. vivax*, and that of humans) using 4,800 compounds as a training set. The ratio of the yields of the HsDHFR and PvDHFR strains were passed to the selection algorithm, together with fingerprints of the remaining 9600 compounds. The results from the first 'cherry-picking' round (compounds selected by active learning and using the hit confirmation assays) (n=96; 12 plates of 8 compounds per plate; 8 replicates of 6 concentrations) were then added to the original data set, and a second cherry-picking round conducted. We used this data to evaluate different approaches to the problem of combining cherry-picking and mass screening data. The approach based on using the mean of replicates multiplied by log(10/conc.) was found to perform best. We

then ran active learning loop was run through three iterations; an initial set of 4,800 compounds was screened (single iteration, 10 μM), and three loops of 96 cherry-picked compounds (8 replicates, at a range of concentrations) were selected. The mean log-weighted cherry-picking data is cycled back into the training set.

5.2. Econometric modelling

A thorough investigation of Eve's QSAR active learning methods, comparing intelligent screening versus standard brute-force screening, would require the analysis of thousands of cycles. We therefore decided to utilize our empirical results from using Eve (in Library screening mode) against the complete set of 14,400 compounds of the Maybridge HitFinder library against dihydrofolate reductase (DHFR) assays from multiple parasitic organisms (see above) - we considered the Johns Hopkin's library to be too small for intelligent screening. The idea is to use these results as an oracle - instead of new physical experiments.

To quantify the utility of intelligent screening we developed an econometric model (Figure 6). In this model the net utility is the cost saving due to not screening compounds, minus the cost due to missing any hits, minus the cumulative cost of the number of active learning cycles performed. Active learning was applied to the seed input data, and predictions made to produce simulated learning curves. The progression of these learning curves was then compared to the base case of standard library screening. For each 96-compound loop the utility equation was applied. Fig. 7 shows the result of one such run involving many cycles of learning, and demonstrates hit enrichment by intelligent screening.

We used the model to investigate a range of costings to determine under what conditions it is economically advantageous compared with performing a standard whole-library screen. Fig. 8 shows that under most conditions it is economically rational to screen intelligently. Assuming that the probability of a compound being a hit is independent of the size of the library i.e. they are independent and identically distributed variables (iid), then the utility gained from intelligent screening is proportional to the size of the library - larger libraries produce larger savings. The iid assumption is reasonable and, in large part, the motivation for the collation of the very large libraries currently used for screening. However it is also conservative, as the difficulties in physically creating libraries means that the probability of an individual compound being a novel structural hit probably decreases with the size of the library, which means that the savings are probably much greater for large libraries. Therefore intelligent screening is more cost-effective with larger libraries, more valuable compounds, and fast cycles of assay screening and testing - this is the standard regime for pharmaceutical screening, suggesting that adoption of intelligent screening is economically rational.

6. Data and Code

To facilitate the confirmation and reuse of Eve's screening assay results we have made this openly available as linked open data (LOD) in Resource Frame Framework (RDF) format (Bizer *et al.*, 2009). A semantic data model of Eve's screening assay results was developed (see Supplementary Information). The root node "assay triple screen" represents the main group of data items used to analyse the results. This root node is linked to the node "Eve" via the relation *ro:has-agent*. The semantics of this association are that Eve initiates and runs the process "assay triple screen". The assay triple screen process has the following inputs (*ro: has-input*): **s**ynthetic yeast strain(s), each has a unique identifier and *ro:has-part* fluorophore and DHFR target;

compound is represented by SMILES code and *sio:has-identifier* compound common name and Maybridge hit finder ID; **p**late is represented by a code and *ro:has-part* well-column and well-row to identify each well. The semantics of these associations are that synthetic yeast strains, compounds, and a plate participate in the assay triple screen process and are present at the beginning of the process. The assay triple screen process has the following outputs (*ro: has-output*): venus, sapphire and cherry initial fluorescence in a well; venus, sapphire and cherry final fluorescence in a well; venus, sapphire and cherry doubling time in a well; venus, sapphire and cherry lagtime2 in a well; venus, sapphire and cherry error code in a well. The semantics of these associations are that initial and final fluorescence, doubling time, lagtime2, and error code measurements were produced by the assay process and are present at the end of the process. Additionally, the relation has-target-origin was introduced to link a target and an organism of origin. We included this relation and other entities that are required to define semantic meaning of Eve data in a small ontology EVE that was specially designed to support the semantic data model of Eve's screening assay results (http://disc.brunel.ac.uk/eve.). The node "DHFR target" is linked via this relation to the host (*Homo sapiens*) and parasites. A target may be drug resistant. This is expressed via the link sio:has-quality. Currently the dataset is deposited at the closed site http://disc.brunel.ac.uk/eve-dataset. On publication Eve dataset will be deposited as LOD (Linked Open Data) (Bizer *et al*., 2009) at the open site Bio2RDF http://bio2rdf.org/eve.

To facilitate the reuse of the code we have placed all the low-level software, and QSAR software used on Github using the GNU General Public License version 3.

6. Discussion and Conclusion

Eve's standardized assays could easily be engineered for other targets/species (e.g. bacteria), or indirect targets (e.g. to drug import or efflux pumps), or more complicated functions (e.g. to screen for drug synergies across multiple targets). In addition the biological realism of the assays could be increased by the incorporation of many more human/parasite genes. The assays could also be modified to be much faster - as using growth as the read-out limits the speed of executing the assay.

The economics of drug development are clearly influenced by many factors (Ioset & Chang, 2011; Leslie, 2011; Braum *et al*., 2010; Pammolli *et al*., 2011) some technical (understanding how to intervene to treat a disease, the difficulty of achieving the intervention, etc.), others societal (safety standards, the price of drug, etc.). Although the costs of drug discovery are substantial, they are relatively small compared to later stages in development. Such arguments tell against increased automation and standardization in drug discovery making much economic difference. However, they fail to take into account the "art of the soluble" (Sir. Peter Medawar). Preventing drug failures in late-stage development is an intrinsically very hard problem, as human biology is very complex. In contrast, we argue that a radical decrease in the cost and increase in the speed of drug discovery could be achieved by the full automation and standardization of procedures. By this we mean a robotic system that once given a target could autonomously develop a standardized assay for that target, screen a compound library using that assay, confirm hit compounds, and identify lead compounds through cycles of quantitative structure activity relationship (QSAR) learning and testing. This could be achieved today: Eve's synthetic biology assays could be automated using existing technology, and chemical synthesis machines exist that could be integrated with Eve (Baumann *et al.*, 2011). Such integration would achieve the goal of a robotic system that could autonomously generate hits for targets, and

radically decrease the cost and increase the speed of drug discovery.

Acknowledgements

This work was supported by grant BB/F008228/1 from the UK Biotechnology & Biological Sciences Research Council and a contract from the European Commission under the FP7 Collaborative Programme, UNICELLSYS, both to SGO and RDK. KDG and JR were supported partially by KU Leuven GOA/08/008 and partially by ERC Starting Grant 240186.

References

1. Arico-Muendel C., *et al*. Antiparasitic activities of novel, orally available fumagillin analogs. *Bioorganic & Medicinal Chemistry Letters* **19**, 5128–5131 (2009)
2. Baumann M., Baxendale I.R., Ley S.V., Martin R.E., Schneider J. Synthesis of a drug-like focused library of trisubstituted pyrrolidines using integrated flow chemistry and batch methods. *ACS Comb Sci* **13**, 405–413 (2011)
3. Besnard J. *et al*. Automated design of ligands to polypharmacological profiles. *Nature* **492**, 215-219 (2012)
4. Benenson Y., Gil B., Ben-Dor U., Adar R., Shapiro E. An autonomous molecular computer for logical control of gene expression. *Nature* **429**, 423-429 (2004)
5. Bernal J.D. *Science in History*. Watts. (1969)
6. Bilsland E. *et al*. Functional expression of parasite drug targets and their human orthologs in yeast *PLoS Negl Trop Dis* **10**, e1320 (2011)
7. Bilsland E. *et al.* Yeast-based automated high-throughput screens to identify antiparasitic lead compounds. *Open Biology* **3**, 120158 (2013)
8. Bizer C., Heath T., Berners-Lee T. Linked data – The story so far. *Intl Semantic Web & Info Systems* **5**, 1-22 (2009)
9. Braun M.M., Farag-El-Massah S., Xu K., Coté T.R. Emergence of orphan drugs in the United States: a quantitative assessment of the first 25 years. *Nature Reviews Drug Discovery* **9**, 519-522 (2010)
10. Chen X., *et al.* Fumagillin and fumarranol interact with *P. falciparum* methionine aminopeptidase 2 and inhibit malaria parasite growth in vitro and in vivo". *Chemistry & Biology* **16**, 193-201 (2009)
11. Chong C.R., Sullivan D.J. New uses for old drugs. *Nature*. **448**, 645-646 (2007)
12. Cohn D.A., Ghahramani Z., Jordan M.I. Active learning with statistical models. *J. Artificial Intell Res* **4**, 129-145 (1996)
13. De Grave K., Ramon J., De Raedt L. Active learning for high throughput screening. *Lecture Notes in Computer Science* **5255**, 185-196, Springer (2008)
14. Deutsch, D. Quantum theory, the Church-Turing principle and the universal quantum computer. *Proceedings of the Royal* Society, *Series A*, **400**, 97-117 (1985)

15. Gad S.C. *Drug Discovery Handbook*. Wiley. (2005)

16. Ioset J.R., Chang S. Drugs for neglected diseases initiative model of drug development for neglected diseases: current status and future challenges. *Future Med Chem* **3**, 1361-1371 (2011)

17. King R.D. *et al*. Functional genomic hypothesis generation and experimentation by a robot scientist *Nature* **427**, 247-252 (2004)

18. King R.D. *et al*. The automation of Science *Science* **324**, 85-89 (2009)

19. Leartsakulpanich U., *et al*. Molecular characterization of dihydrofolate reductase in relation to antifolate resistance in *Plasmodium vivax*. *Molecular and Biochemical Parasitology*. **119**, 63-73 (2002)

20. Leslie M. Drug developers finally take aim at a neglected disease. *Science* **333**, 933-935 (2011)

21. Lounkine E. *et al*. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**, 361-367 (2012)

22. Martin Y.C. Quantitative drug design: A critical introduction, second edition. CRC Press. (2010)

23. O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* 3: 33. (2011)

24. Pammolli F., Magazzini L., Riccaboni M. The productivity crisis in pharmaceutical R&D. *Nature Reviews Drug Discovery* **10**, 428-437 (2011)

25. Rasmussen C.E., Williams C.K.I. Gaussian processes for machine learning. The MIT Press. (2006)

26. Schneider, G. Virtual screening: an endless staircase? *Nature Reviews Drug Discovery* **9**, 273-276 (2010)

27. Smith A.M., Ammar R., Nislow C., Giaever G. A survey of yeast genomic assays for drug and target discovery. *Pharmacol Ther* **127**, 156–164 (2010)

28. Sparkes A *et al*. Towards robot scientists for autonomous scientific discovery *Autom Exp* **2**, 1-22 (2010)

29. Vogel, H.G. *Drug Discovery and Evaluation: Pharmacological Assays*. Springer. (2002)

30. Yuthavong *et al*. Malarial dihydrofolate reductase as a paradigm for drug development against a resistance-compromised target *Proc Natl Acad Sci USA* **109**, 16823–16828 (2012)

31. Zhang P. *et al*. Angiogenesis inhibitors specific for methionine aminopeptidase 2 as drugs for malaria and leishmaniasis. *J. Biomed Sci* **9**, 34-40 (2002)

Disease	**Species**	**Enzyme**	**Drug resistant**	**Libraries**
malaria	*P. falciparum*	DHFR	no	May
malaria	*P. falciparum*	DHFR	yes	May
malaria	*P. falciparum*	DHFR	no	May
malaria	*P. vivax*	DHFR	no	May
malaria	*P. vivax*	DHFR	yes	May
malaria	*P. vivax*	DHFR	no	May
malaria	*P. vivax*	PGK	no	May
malaria	*P. vivax*	NMT	no	May
Chagas	*T. cruzi*	DHFR	no	May
Chagas	*T. cruzi*	PGK	no	May
Chagas	*T. cruzi*	NMT	no	May
African sleeping sickness	*T. bruci*	DHFR	no	May
African sleeping sickness	*T. bruci*	PGK	no	May
African sleeping sickness	*T. bruci*	NMT	no	May
schistosomiasis	*S. mansoni*	DHFR	no	May
schistosomiasis	*S. mansoni*	PGK	no	May
schistosomiasis	*S. mansoni*	NMT	no	May
leishmaniasis	*L. major*	DHFR	no	May
bacterial infection	*S. aureus*	DHFR	no	May

Table 1. The targets (disease/ species/ protein/ drug-resistant) and libraries screened (May = Maybridge Hitfinder, JH –Johns Hopkins University Clinical Compound Library)

Figure Legends

Figure 1. **Early-stage Drug Design**. The contribution of standardized synthetic biology assays and Eve to a cheaper faster drug discovery pipeline.

Figure 2a. A diagram of Eve. Showing the location of the main instruments.

Figure 2b. A photo of Eve. Eve has been designed to be flexible in terms of the biological assays that it can perform, and is physically capable of screening at a moderately high throughput rate.

Figure 3. The form of the standardized assays. From biological knowledge a specific objective is determined; this is compiled into a propositional logic function, and synthetic biology is used to engineer a set of yeast strains that compute the function.

Figure 4. The structure of TNP-470

Figure 5. An Eve hit confirmation run with four replicates. TNP-470 dose response curves for: y^{Hs}DHFRp (red), y^{Pf}DHFRp (green), and y^{Pv}DHFRp (blue). Normalized growth is calculated by comparison to in-plate negative controls.

Figure 6. Modelling the economics of drug discovery. The econometric model of the differential utility of intelligent screening versus mass screening with hit confirmation.

Figure 7. Intelligent v Random Screening. An example simulation run of intelligent screening: cycles of QSAR learning/testing from a compound library. The data are taken for a screen of the Maybridge Hitfinder library against the *P. vivax* DHFR as target (Supplementary material). Intelligent screening is red, and standard brute-force black. The differential utility of intelligent screening is shown in blue, and the dotted line indicates its maximum. It can be seen that it is cost-optimal to screen between a third and a half of Eve's small library, with a larger library the screened proportion would be expected to be smaller. Similar diagrams for the other targets can be found in the Supplementary material.

Figure 8. Summary of Utility Modelling. Diagram of the maximum utility of intelligent screening taken from a systematic scan of different costs/utilities in the econometric model (A), using the screening results in (B). To make these results comprehensible, we project them down into a 3-dimensional graph and combine cost/utilities: Time-ratio = Tc/Tm, and Cost-ratio = Uh/Cc. This indicates that intelligent screening is generally rational (there is little area of

negative utility), and that a high time-ratio (fast screening) and low cost-ratio (valuable library compounds) are most favorable.